Linear Algebra and its Applications 436 (2012) 3109–3119

Contents lists available at SciVerse ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa

Structured backward errors for generalized saddle point systems ☆

Xiao Shan Chen [a], Wen Li [a,*], Xiaojun Chen [b], Jun Liu [a]

[a] School of Mathematical Sciences, South China Normal University, Guangzhou 510631, People's Republic of China
[b] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong

ARTICLE INFO

Article history:
Received 13 August 2011
Accepted 3 October 2011
Available online 8 November 2011

Submitted by C.-K. Li

AMS classification:
65F10

Keywords:
Generalized saddle point system
Structured backward error
Frobenius norm

ABSTRACT

In this paper we investigate structured backward errors for three kinds of generalized saddle point systems where the matrix is not symmetric and its (2, 2) block is not zero and has perturbations. Computable formula of backward errors are derived. The expressions are useful for testing the stability of practical algorithms.

1. Introduction

The purpose of this paper is to consider the structured backward error of the following block two-by-two linear system:

$$\begin{bmatrix} A & E^T \\ F & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b \\ c \end{bmatrix}, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$, $E, F \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{m\times m}$ and E^T stands for the transpose of E. Backward error analysis can answer how close the problem that is actually solved is to the one we want to solve

☆ This work was supported by the Natural Science Foundations of Guangdong Province (06025061, 9151063101000021, S2011040003243), the National Natural Science Foundations of China (10671077, 10971075), Research Fund for the Doctoral Program of Higher Education of China (20104407110001), the Opening Project of Guangdong Province Key Laboratory of Computational Science of Sun Yat-sen University (201106005).

* Corresponding author.
E-mail address: ma.wenli@yahoo.com.cn (W. Li)

0024-3795/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2011.10.012

and reveals the stability of a numerical method [4,12]. It is obvious that any linear system can be presented in the block form (1.1). When A is symmetric and $E = F$, $C = 0$ (zero matrix), (1.1) is called a Karush–Kuhn–Tucker (KKT) system [2,20]. The linear system (1.1) is called a generalized saddle point problem if the blocks A, E, F and C satisfy some special structures, for example A is symmetric, $F \neq E$ or $C \neq 0$. Generalized saddle point systems arises from many important problems in optimization and numerical differential equations [1–3,7,13,15].

To simplify our discussion, we can write (1.1) as

$$Mz = p. \tag{1.2}$$

For a computed solution $\tilde{z}$ the normwise backward error $\eta(\tilde{z})$ (see Rigal and Gaches' definition [16,19,20]) is defined by

$$\eta(\tilde{z}) = \min_{(\Delta M, \Delta p)\in\mathcal{G}} \left\| \left(\frac{\|\Delta M\|_F}{\|M\|_F}, \frac{\|\Delta p\|_2}{\|p\|_2} \right) \right\|_2,$$

where $\mathcal{G}$ is defined by

$$\mathcal{G} = \{(\Delta M, \Delta p) : (M + \Delta M)\tilde{z} = p + \Delta p\}.$$

$\eta(\tilde{z})$ can be expressed by [19,20]

$$\eta(\tilde{z}) = \frac{\|p - M\tilde{z}\|_2}{\sqrt{\|M\|_F^2 \|\tilde{z}\|_2^2 + \|p\|_2^2}}, \tag{1.3}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\|\cdot\|_2$ is the Euclidean norm.

A small $\eta(\tilde{z})$ means that the computed solution $\tilde{z}$ is the exact solution of a slightly perturbed linear system $\tilde{M}\tilde{z} = \tilde{p}$, which implies that the backward error $\eta(\tilde{z})$ can be applied to test the stability of the algorithms for solving the linear system (1.2).

However the coefficient matrix M of the system (1.2) has a more special structure, a natural requirement is that the perturbed matrix $\tilde{M}$ also has the same one as M. For this case, Bunch [4] defined that an algorithm for solving Eq. (1.2) is strongly stable if the coefficient matrix M and the perturbed matrix $\tilde{M}$ have the same structure. When M in (1.2) is a symmetric matrix, Bunch et al. [5] gave the stability analysis for solving symmetric linear systems. When $A^T = A$, $E = F$ and $C = 0$ in (1.1), Sun [20] defined the structured backward error for a computed solution and obtained its explicit expression. Based on the Sun's technique [20], some authors gave explicit expressions of the structured backward errors for the system (1.1) with the following structures: (i) $A = I$ (identity matrix), $E = F$ and $C = 0$ [14]; (ii) $A^T \neq A$, $E = F$ and $C = 0$ [22]; (iii) $A^T \neq A$, $E \neq F$ and $C = 0$ [6].

In this paper, we consider the linear system (1.1) where the (2, 2) block is not a zero matrix and there are perturbations in the (2, 2) block. The aim of this paper is to investigate structured backward errors for the following three cases: (i) $E = F$, $C = C^T$; (ii) $A = A^T$, $C = C^T$; (iii) $E = F$.

Let θ_1, θ_2, θ_3, θ_4, λ and μ be positive parameters. Taking into consideration the special block structure of (1.1), we define the normwise structured backward errors $\eta_{S_1}(\tilde{z})$, $\eta_{S_2}(\tilde{z})$ and $\eta_{S_3}(\tilde{z})$ with respect to a computed solution $\tilde{z} = \left(\tilde{x}^T, \tilde{y}^T\right)^T$ for the above three cases, respectively

$$\eta_{S_1}(\tilde{z}) = \min_{(\Delta A, \Delta E, \Delta C, \Delta b, \Delta c)\in\mathcal{E}_1} \|(\theta_1\|\Delta A\|_F, \theta_2\|\Delta E\|_F, \theta_4\|\Delta C\|_F, \lambda\|\Delta b\|_2, \mu\|\Delta c\|_2)\|_2, \tag{1.4}$$

$$\eta_{S_2}(\tilde{z}) = \min_{\left(\begin{smallmatrix}\Delta A, \Delta E, \Delta F,\\ \Delta C, \Delta b, \Delta c\end{smallmatrix}\right)\in\mathcal{E}_2} \|(\theta_1\|\Delta A\|_F, \theta_2\|\Delta E\|_F, \theta_3\|\Delta F\|_F, \theta_4\|\Delta C\|_F, \lambda\|\Delta b\|_2, \mu\|\Delta c\|_2)\|_2 \tag{1.5}$$

and

$$\eta_{S_3}(\tilde{z}) = \min_{\begin{pmatrix} \Delta A,\, \Delta E, \\ \Delta C,\, \Delta b,\, \Delta c \end{pmatrix} \in \mathcal{E}_3} \|(\theta_1\|\Delta A\|_F, \theta_2\|\Delta E\|_F, \theta_4\|\Delta C\|_F, \lambda\|\Delta b\|_2, \mu\|\Delta c\|_2)\|_2, \tag{1.6}$$

where the sets $\mathcal{E}_1$, $\mathcal{E}_2$ and $\mathcal{E}_3$ are defined by

$$\mathcal{E}_1 = \left\{ (\Delta A, \Delta E, \Delta C, \Delta b, \Delta c) : \begin{array}{c} \begin{bmatrix} A+\Delta A & (E+\Delta E)^T \\ E+\Delta E & C+\Delta C \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \\ = \begin{bmatrix} b+\Delta b \\ c+\Delta c \end{bmatrix}, \ \Delta C^T = \Delta C \end{array} \right\}, \tag{1.7}$$

$$\mathcal{E}_2 = \left\{ \begin{pmatrix} \Delta A, \Delta E, \Delta F, \\ \Delta C, \Delta b, \Delta c \end{pmatrix} : \begin{array}{c} \begin{bmatrix} A+\Delta A & (E+\Delta E)^T \\ F+\Delta F & C+\Delta C \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \\ = \begin{bmatrix} b+\Delta b \\ c+\Delta c \end{bmatrix}, \ \begin{array}{c} \Delta A^T = \Delta A \\ \Delta C^T = \Delta C \end{array} \end{array} \right\} \tag{1.8}$$

and

$$\mathcal{E}_3 = \left\{ \begin{pmatrix} \Delta A, \Delta E, \\ \Delta C, \Delta b, \Delta c \end{pmatrix} : \begin{array}{c} \begin{bmatrix} A+\Delta A & (E+\Delta E)^T \\ E+\Delta E & C+\Delta C \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{y} \end{bmatrix} \\ = \begin{bmatrix} b+\Delta b \\ c+\Delta c \end{bmatrix}, \end{array} \right\}, \tag{1.9}$$

respectively.

Remark 1.1. If we take $\theta_1 = \theta_2 = \theta_3 = \theta_4 = \lambda = \mu = 1$, then the structured backward errors in (1.4)–(1.6) are called the absolute ones. If we take $\frac{1}{\theta_1} = \|A\|_F \neq 0$, $\frac{1}{\theta_2} = \|E\|_F \neq 0$, $\frac{1}{\theta_3} = \|F\|_F \neq 0$, $\frac{1}{\theta_4} = \|C\|_F \neq 0$, $\frac{1}{\lambda} = \|b\|_2 \neq 0$, and $\frac{1}{\mu} = \|c\|_2 \neq 0$, the above backward errors are called the relative ones.

Let $\tilde{z} = (\tilde{x}^T, \tilde{y}^T)^T$ be a computed solution given by some algorithms. By the definitions of $\eta_{S_i}(\tilde{z})$, $i = 1, 2, 3$, if $\eta_{S_i}(\tilde{z})$ is small, then the computed solution $\tilde{z}$ is a solution to a nearby system, which has the same structure as its original one. For this case, the algorithm for solving the corresponding structured linear system is strongly stable. The rest of this paper is organized as follows. In Sections 2–4, we give the explicit expressions of $\eta_{S_1}(\tilde{z})$, $\eta_{S_2}(\tilde{z})$ and $\eta_{S_3}(\tilde{z})$, respectively. In Section 5, a numerical example is given to illustrate strongly stability of the GEPP and GMRES algorithms for solving generalized saddle point systems. In Section 6, concluding remarks are given.

Throughout this paper, we adopt the following notations. $A^\dagger$ stands for the Moore–Penrose inverse of A. $P_A = AA^\dagger$ is the orthogonal projection onto the column space of A and $P_A^\perp = I - P_A$. For $A = (a_1, \cdots, a_n) = (a_{ij}) \in \mathbb{R}^{n\times n}$ and a matrix B, $A \otimes B = (a_{ij}B)$ is a Kronecker product, and $\mathrm{vec}(A)$ is a vector defined by $\mathrm{vec}(A) = (a_1^T, \cdots, a_n^T)^T$. For $A \in \mathbb{R}^{m\times n}$, we have

$$\mathrm{vec}(A^T) = \Pi \mathrm{vec}(A),$$

where Π is the vec-permutation matrix which can be expressed by

$$\Pi = \sum_{k=1}^{m}\sum_{l=1}^{n} e_k^{(m)} e_l^{(n)^T} \otimes e_l^{(n)} e_k^{(m)^T},$$

in which $e_k^{(m)}$ denotes the kth column of an $m \times m$ identity matrix I_m.

2. Expression of $\eta_{S_1}(\tilde{z})$

In this section we give the explicit expression of the backward error $\eta_{S_1}(\tilde{z})$. In order to prove our results, the following lemmas are useful.

Lemma 2.1 [18]. *Let $f \in \mathbb{R}^m$ and $g \in \mathbb{R}^n$ be given. Define*

$$\mathscr{X} = \{X \in \mathbb{R}^{n\times m} : \quad Xf = g\}.$$

Then $\mathscr{X} \neq \emptyset$ if and only if f and g satisfy $gf^{\dagger}f = g$, and in the case of $\mathscr{X} \neq \emptyset$, and any $X \in \mathscr{X}$ can be expressed by

$$X = gf^{\dagger} + Z(I - ff^{\dagger}), \quad Z \in \mathbb{R}^{n\times m}.$$

Lemma 2.2 [18]. *Let $b, c \in \mathbb{R}^n$ be given. Define*

$$\mathscr{H} = \left\{H \in \mathbb{R}^{n\times n} : \quad H^T = H,\ Hb = c\right\}.$$

Then $\mathscr{H} \neq \emptyset$ if and only if b and c satisfy $cb^{\dagger}b = c$, and in the case of $\mathscr{H} \neq \emptyset$, any $H \in \mathscr{H}$ can be expressed by

$$H = cb^{\dagger} + \left(cb^{\dagger}\right)^T - \left(b^{\dagger}\right)^T c^T bb^{\dagger} - (I - bb^{\dagger})T(I - bb^{\dagger}),$$

where $T \in \mathbb{R}^{n\times n}$ and $T^T = T$.

We obtain the following expression of $\eta_{S_1}(\tilde{z})$.

Theorem 2.3. *Let $\tilde{z} = \left(\tilde{x}^T, \tilde{y}^T\right)^T$ be a computed solution of the system (1.1) with $E = F$, $C^T = C$ and $\tilde{y} \neq 0$. Then we have*

$$\eta_{S_1}(\tilde{z}) = \left\|P_K^{\perp} d\right\|_2, \tag{2.1}$$

where

$$K = \begin{pmatrix} \mu I & 0 \\ 0 & \theta_2 I \\ 0 & -\frac{1}{\gamma}\left(\tilde{y}^T \otimes I\right)\Pi \\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^T & -\frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{x}^T \otimes \tilde{y}^T \\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}\left(I - \tilde{y}\tilde{y}^{\dagger}\right) & -\frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}\tilde{x}^T \otimes \left(I - \tilde{y}\tilde{y}^{\dagger}\right) \end{pmatrix} \in \mathbb{R}^{l\times(mn+m)}, \tag{2.2}$$

$$d = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{\gamma}r_1 \\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^T r_2 \\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}\left(I - \tilde{y}\tilde{y}^{\dagger}\right) r_2 \end{pmatrix} \in \mathbb{R}^l, \quad l = mn + 2m + n + 1. \tag{2.3}$$

and

$$r_1 = b - A\tilde{x} - E^T\tilde{y}, \quad r_2 = c - E\tilde{x} - C\tilde{y}, \quad \gamma = \sqrt{\frac{1}{\theta_1^2}\|\tilde{x}\|_2^2 + \frac{1}{\lambda^2}}. \tag{2.4}$$

Proof. By (1.7), $(\Delta A, \Delta E, \Delta C, \Delta b, \Delta c) \in \mathcal{E}_1$ if and only if ΔA, ΔE, ΔC, Δb and Δc satisfy

$$\Delta A\tilde{x} + \Delta E^T\tilde{y} - \Delta b = r_1, \tag{2.5}$$

$$\Delta E\tilde{x} + \Delta C\tilde{y} - \Delta c = r_2, \quad \Delta C^T = \Delta C, \tag{2.6}$$

where r_1, r_2 are defined by (2.4). The equation (2.5) can be rewritten as

$$\left[\ \theta_1\Delta A,\ -\lambda\Delta b\ \right]\begin{bmatrix} \frac{1}{\theta_1}\tilde{x} \\ \frac{1}{\lambda} \end{bmatrix} = r_1 - \Delta E^T\tilde{y}. \tag{2.7}$$

Let

$$z = \begin{bmatrix} \frac{1}{\theta_1}\tilde{x} \\ \frac{1}{\lambda} \end{bmatrix}. \tag{2.8}$$

Since $z \neq 0$, applying Lemma 2.1 to Eq. (2.7) gives that the equation is solvable and any solution $[\theta_1\Delta A,\ -\lambda\Delta b]$ can be expressed by

$$\left[\ \theta_1\Delta A,\ -\lambda\Delta b\ \right] = \left(r_1 - \Delta E^T\tilde{y}\right)z^\dagger + Z\left(I - zz^\dagger\right), \tag{2.9}$$

where $Z \in \mathbb{R}^{n\times(n+1)}$. Let $v_1 = z/\gamma$ and choose $V_2 \in \mathbb{R}^{(n+1)\times n}$ so that $V = (v_1,\ V_2)$ is an $(n+1)\times(n+1)$ orthogonal matrix, where γ is defined by (2.4). Then by (2.9) we have

$$\left[\ \theta_1\Delta A,\ -\lambda\Delta b\ \right](v_1,\ V_2) = \left[\ \frac{1}{\gamma}\left(r_1 - \Delta E^T\tilde{y}\right),\ ZV_2\ \right].$$

Hence

$$\theta_1^2\|\Delta A\|_F^2 + \lambda^2\|\Delta b\|_2^2 = \frac{1}{\gamma^2}\|r_1 - \Delta E^T\tilde{y}\|_2^2 + \|ZV_2\|_F^2 \equiv \Phi(\Delta E, Z). \tag{2.10}$$

Eq. (2.6) can be written as

$$(\theta_4\Delta C)\left(\frac{1}{\theta_4}\tilde{y}\right) = r_2 + \Delta c - \Delta E\tilde{x} \equiv w, \quad \Delta C^T = \Delta C. \tag{2.11}$$

Since $\tilde{y} \neq 0$, by Lemma 2.2, Eq. (2.11) is solvable and any solution $\theta_4\Delta C$ can be expressed as

$$\begin{aligned}\theta_4\Delta C = {} & \theta_4 w\tilde{y}^\dagger + \theta_4\left(w\tilde{y}^\dagger\right)^T - \theta_4\left(\tilde{y}^\dagger\right)^T w^T\tilde{y}\tilde{y}^\dagger \\ & + \left(I - \tilde{y}\tilde{y}^\dagger\right)T\left(I - \tilde{y}\tilde{y}^\dagger\right),\end{aligned} \tag{2.12}$$

where $T^T = T \in \mathbb{R}^{m\times m}$. Let $u_1 = \tilde{y}/\|\tilde{y}\|_2$, and choose $U_2 \in \mathbb{R}^{m\times(m-1)}$ so that the matrix $U = (u_1, U_2)$ is an $m \times m$ orthogonal matrix. Then we get from (2.12)

$$U^T(\theta_4\Delta C)U = \begin{bmatrix} \frac{2\theta_4}{\|\tilde{y}\|_2}u_1^Tw & \frac{\theta_4}{\|\tilde{y}\|_2}w^TU_2 \\ \frac{\theta_4}{\|\tilde{y}\|_2}U_2^Tw & U_2^TTU_2 \end{bmatrix}. \tag{2.13}$$

By $\left\|U_2^Tw\right\|_F = \left\|(U_2U_2^T)w\right\|_F = \left\|\left(I - \tilde{y}\tilde{y}^\dagger\right)w\right\|_F$ and (2.13), we obtain

$$\begin{aligned}\|\theta_4\Delta C\|_F^2 &= \frac{4\theta_4^2}{\|\tilde{y}\|_2^4}\left(\tilde{y}^T w\right)^2 + 2\frac{\theta_4^2}{\|\tilde{y}\|_2^2}\left\|\left(I-\tilde{y}\tilde{y}^\dagger\right)w\right\|_F^2 + \left\|U_2^T T U_2\right\|_F^2 \\ &\equiv \Psi(\Delta E,\ \Delta c,\ T).\end{aligned} \tag{2.14}$$

Hence from the definition (1.4) of $\eta_{S_1}(\tilde{z})$, (2.10) and (2.14), we can get

$$\begin{aligned}\left[\eta_{S_1}(\tilde{z})\right]^2 &= \min_{\substack{\Delta E\in\mathbb{R}^{m\times n},\ \Delta c\in\mathbb{R}^m \\ Z\in\mathbb{R}^{n\times(n+1)},\ T\in\mathbb{R}^{m\times m},\ T^T=T}} \left[\theta_2^2\|\Delta E\|_F^2+\mu^2\|\Delta c\|_2^2+\Phi(\Delta E,Z)+\Psi(\Delta E,\Delta c,T)\right] \\ &= \min_{\Delta E\in\mathbb{R}^{m\times n},\ \Delta c\in\mathbb{R}^m}\left[\theta_2^2\|\Delta E\|_F^2+\mu^2\|\Delta c\|_2^2+\frac{1}{\gamma^2}\left\|r_1-\Delta E^T\tilde{y}\right\|_2^2\right. \\ &\quad \left.+\frac{4\theta_4^2}{\|\tilde{y}\|_2^4}\left(\tilde{y}^T w\right)^2+\frac{2\theta_4^2}{\|\tilde{y}\|_2^2}\left\|\left(I-\tilde{y}\tilde{y}^\dagger\right)w\right\|_F^2\right].\end{aligned} \tag{2.15}$$

Using the Kronecker product $\otimes$ (e. g. see [10]), we have

$$\begin{aligned}&r_1-\Delta E^T\tilde{y} = r_1-[\tilde{y}^T\otimes I]\Pi\mathrm{vec}(\Delta E),\\ &\tilde{y}^T w=\tilde{y}^T r_2+\tilde{y}^T\Delta c-\left(\tilde{x}^T\otimes\tilde{y}^T\right)\mathrm{vec}(\Delta E),\\ &\left(I-\tilde{y}\tilde{y}^\dagger\right)w=\left(I-\tilde{y}\tilde{y}^\dagger\right)r_2+\left(I-\tilde{y}\tilde{y}^\dagger\right)\Delta c-\left[\tilde{x}^T\otimes\left(I-\tilde{y}\tilde{y}^\dagger\right)\right]\mathrm{vec}(\Delta E).\end{aligned} \tag{2.16}$$

By (2.16), we know that (2.15) can be written as the following form:

$$\left[\eta_{S_1}(\tilde{z})\right]^2 = \min_{\Delta E\in\mathbb{R}^{m\times n},\ \Delta c\in\mathbb{R}^m}\left\|d+K\begin{pmatrix}\Delta c\\ \mathrm{vec}(\Delta E)\end{pmatrix}\right\|_2^2, \tag{2.17}$$

where K and d are defined by (2.2) and (2.3), respectively. Solving the least square problem (2.17) (e.g., see [11]) gives the expression (2.1). □

Remark 2.1. It is assumed in Theorem 2.3 that $\tilde{y}\neq 0$. If $\tilde{x}\neq 0$ and $\tilde{y}=0$, then it is easy to obtain

$$\eta_{S_1}(\tilde{z}) = \sqrt{\frac{\theta_1^2\lambda^2\|b-A\tilde{x}\|_2^2}{\lambda^2\|\tilde{x}\|_2^2+\theta_1^2}+\frac{\theta_2^2\mu^2\|c-B\tilde{x}\|_2^2}{\mu^2\|\tilde{x}\|_2^2+\theta_2^2}}.$$

3. Expression of $\eta_{S_2}(\tilde{z})$

In this section, we give the following explicit expression of $\eta_{S_2}(\tilde{z})$ with respect to a computable solution $\tilde{z}=(\tilde{x}^T,\tilde{y}^T)^T$ to the generalized saddle point system (1.1) with $A^T=A$, $C^T=C$.

Theorem 3.1. *Let $\tilde{z}=\left(\tilde{x}^T,\tilde{y}^T\right)^T$ be a computed solution of the system (1.1) with $A^T=A$, $C^T=C$, and $\tilde{x}\neq 0$, $\tilde{y}\neq 0$. Then we have*

$$\eta_{S_2}(\tilde{z}) = \sqrt{\left\|P_{K_1}^{\perp}d_1\right\|_2^2+\left\|P_{K_2}^{\perp}d_2\right\|_2^2}, \tag{3.1}$$

where

$$K_1 = \begin{bmatrix} \lambda I & 0 \\ 0 & \theta_2 I \\ \frac{2\theta_1}{\|\tilde{x}\|_2^2}\tilde{x}^T & -\frac{2\theta_1}{\|\tilde{x}\|_2^2}(\tilde{y}^T \otimes \tilde{x}^T)\Pi \\ \frac{\sqrt{2}\theta_1}{\|\tilde{x}\|_2}(I - \tilde{x}\tilde{x}^\dagger) & -\frac{\sqrt{2}\theta_1}{\|\tilde{x}\|_2}(\tilde{y}^T \otimes (I - \tilde{x}\tilde{x}^\dagger))\Pi \end{bmatrix} \in \mathbb{R}^{l_1 \times (mn+n)}, \tag{3.2}$$

$$K_2 = \begin{bmatrix} \mu I & 0 \\ 0 & \theta_3 I \\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^T & -\frac{2\theta_4}{\|\tilde{y}\|_2^2}(\tilde{x}^T \otimes \tilde{y}^T) \\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}(I - \tilde{y}\tilde{y}^\dagger) & -\frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}(\tilde{x}^T \otimes (I - \tilde{y}\tilde{y}^\dagger)) \end{bmatrix} \in \mathbb{R}^{l_2 \times (mn+n)}, \tag{3.3}$$

$$d_1 = \begin{bmatrix} 0 \\ 0 \\ \frac{2\theta_1}{\|\tilde{x}\|_2^2}\tilde{x}^T r_1 \\ \frac{\sqrt{2}\theta_1}{\|\tilde{x}\|_2}(I - \tilde{x}\tilde{x}^\dagger) r_1 \end{bmatrix} \in \mathbb{R}^{l_1}, \quad d_2 = \begin{bmatrix} 0 \\ 0 \\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^T r_2 \\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}(I - \tilde{y}\tilde{y}^\dagger) r_2 \end{bmatrix} \in \mathbb{R}^{l_2}, \tag{3.4}$$

$$l_1 = mn + 2n + 1, \quad l_2 = mn + 2m + 1,$$

and

$$r_1 = b - A\tilde{x} - E^T\tilde{y}, \quad r_2 = c - F\tilde{x} - C\tilde{y}. \tag{3.5}$$

Proof. From (1.8), $(\Delta A, \Delta E, \Delta F, \Delta C, \Delta b, \Delta c) \in \mathcal{E}_2$ if and only if ΔA, ΔE, ΔF, ΔC, Δb and Δc satisfy

$$(\theta_1 \Delta A)\left(\frac{1}{\theta_1}\tilde{x}\right) = r_1 - \Delta E^T\tilde{y} + \Delta b \equiv w_1, \quad \Delta A^T = \Delta A \tag{3.6}$$

and

$$(\theta_4 \Delta C)\left(\frac{1}{\theta_4}\tilde{y}\right) = r_2 - \Delta F\tilde{x} + \Delta c \equiv w_2, \quad \Delta C^T = \Delta C, \tag{3.7}$$

where r_1 and r_2 are defined by (3.5). Since $\tilde{x} \neq 0$, applying Lemma 2.2 to (3.6) gives

$$\begin{aligned} \|\theta_1^2 \Delta A\|_F^2 &= \frac{4\theta_1^2}{\|\tilde{x}\|_2^4}\left(\tilde{x}^T w_1\right)^2 + \frac{2\theta_1^2}{\|\tilde{x}\|_2^2}\left\|(I - \tilde{x}\tilde{x}^\dagger)w_1\right\|_2^2 + \left\|V_2^T T_1 V_2\right\|_F^2 \\ &\equiv \Phi_1(\Delta E, \Delta b, T_1), \end{aligned} \tag{3.8}$$

where $T_1^T = T_1 \in \mathbb{R}^{n\times n}$ and $V_2 \in \mathbb{R}^{n\times(n-1)}$ is chosen such that $(\tilde{x}/\|\tilde{x}\|_2, V_2)$ is an $n \times n$ orthogonal matrix. Similarly, by $\tilde{y} \neq 0$, applying Lemma 2.2 to (3.7) gives

$$\begin{aligned} \|\theta_4^2 \Delta C\|_F^2 &= \frac{4\theta_4^2}{\|\tilde{y}\|_2^4}\left(\tilde{y}^T w_2\right)^2 + \frac{2\theta_4^2}{\|\tilde{y}\|_2^2}\left\|(I - \tilde{y}\tilde{y}^\dagger)w_2\right\|_2^2 + \left\|W_2^T T_2 W_2\right\|_F^2 \\ &\equiv \Phi_2(\Delta F, \Delta c, T_2), \end{aligned} \tag{3.9}$$

where $T_2^T = T_2 \in \mathbb{R}^{m\times m}$ and $(\tilde{y}/\|\tilde{y}\|_2, W_2)$ is an $m \times m$ orthogonal matrix. From definitions (1.5), (3.8) and (3.9), it is easy to see that

$$
\begin{aligned}
[\eta_{S_2}(\tilde{z})]^2 = & \min_{\substack{\Delta E\in\mathbb{R}^{m\times n},\,\Delta b\in\mathbb{R}^n,\,T_1^T=T_1\in\mathbb{R}^{n\times n}\\ \Delta F\in\mathbb{R}^{m\times n},\,\Delta c\in\mathbb{R}^n,\,T_2^T=T_2\in\mathbb{R}^{m\times m}}} \Big[\theta_2^2\|\Delta E\|_F^2+\lambda^2\|\Delta b\|_2^2+\Phi_1(\Delta E,\Delta b,T_1)\\
& +\theta_3^2\|\Delta F\|_F^2+\mu^2\|\Delta c\|_2^2+\Phi_2(\Delta F,\Delta c,T_2)\Big]\\
= & \min_{\Delta E\in\mathbb{R}^{m\times n},\Delta b\in\mathbb{R}^n}\left[\theta_2^2\|\Delta E\|_F^2+\lambda^2\|\Delta b\|_2^2+\frac{4\theta_1^2}{\|\tilde{x}\|_2^4}\left(\tilde{x}^Tw_1\right)^2+\frac{2\theta_1^2}{\|\tilde{x}\|_2^2}\left\|(I-\tilde{x}\tilde{x}^\dagger)w_1\right\|_2^2\right]\\
& +\min_{\Delta F\in\mathbb{R}^{m\times n},\Delta c\in\mathbb{R}^n}\left[\theta_3^2\|\Delta F\|_F^2+\mu^2\|\Delta c\|_2^2+\frac{4\theta_4^2}{\|\tilde{y}\|_2^4}\left(\tilde{y}^Tw_2\right)^2+\frac{2\theta_4^2}{\|\tilde{y}\|_2^2}\left\|(I-\tilde{y}\tilde{y}^\dagger)w_2\right\|_2^2\right].
\end{aligned}
\tag{3.10}
$$

By (3.6) and (3.7) we have

$$\tilde{x}^Tw_1=\tilde{x}^Tr_1-\left(\tilde{y}^T\otimes\tilde{x}^T\right)\Pi\mathrm{vec}(\Delta E)+\tilde{x}^T\Delta b, \tag{3.11}$$

$$(I-\tilde{x}\tilde{x}^\dagger)w_1=(I-\tilde{x}\tilde{x}^\dagger)r_1-\left[\tilde{y}^T\otimes(I-\tilde{x}\tilde{x}^\dagger)\right]\Pi\mathrm{vec}(\Delta E)+(I-\tilde{x}\tilde{x}^\dagger)\Delta b, \tag{3.12}$$

$$\tilde{y}^Tw_2=\tilde{y}^Tr_2-\left(\tilde{x}^T\otimes\tilde{y}^T\right)\mathrm{vec}(\Delta F)+\tilde{y}^T\Delta c \tag{3.13}$$

and

$$(I-\tilde{y}\tilde{y}^\dagger)w_2=(I-\tilde{y}\tilde{y}^\dagger)r_2-\left[\tilde{x}^T\otimes(I-\tilde{y}\tilde{y}^\dagger)\right]\mathrm{vec}(\Delta F)+(I-\tilde{y}\tilde{y}^\dagger)\Delta c. \tag{3.14}$$

By (3.11)–(3.14), (3.10) can be written as follows:

$$
\begin{aligned}
[\eta_{S_2}(\tilde{z})]^2 = & \min_{\Delta E\in\mathbb{R}^{m\times n},\,\Delta b\in\mathbb{R}^n}\left\|d_1+K_1\begin{pmatrix}\Delta b\\ \mathrm{vec}(\Delta E)\end{pmatrix}\right\|_2^2\\
& +\min_{\Delta F\in\mathbb{R}^{m\times n},\,\Delta c\in\mathbb{R}^m}\left\|d_2+K_2\begin{pmatrix}\Delta c\\ \mathrm{vec}(\Delta F)\end{pmatrix}\right\|_2^2
\end{aligned}
\tag{3.15}
$$

where K_1, K_2, d_1 and d_2 are defined by (3.2)–(3.4), respectively. Solving the least square problem (3.15) gives the expression (3.1). □

Remark 3.1. It is assumed in Theorem 3.1 that $\tilde{x}\neq 0$ and $\tilde{y}\neq 0$. If $\tilde{x}=0$ and $\tilde{y}\neq 0$, then we can obtain

$$\eta_{S_2}(\tilde{z})=\sqrt{\frac{\theta_2^2\lambda^2\|b-E^T\tilde{y}\|_2^2}{\lambda^2\|\tilde{y}\|_2^2+\theta_2^2}+\left\|P_{K_3}^{\perp}d_3\right\|_2^2}$$

where

$$d_3=\begin{bmatrix}0\\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^Tr_3\\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}\left(I-\tilde{y}\tilde{y}^\dagger\right)r_3\end{bmatrix}\in\mathbb{R}^{l_3},\quad K_3=\begin{bmatrix}\mu I\\ \frac{2\theta_4}{\|\tilde{y}\|_2^2}\tilde{y}^T\\ \frac{\sqrt{2}\theta_4}{\|\tilde{y}\|_2}\left(I-\tilde{y}\tilde{y}^\dagger\right)\end{bmatrix}\in\mathbb{R}^{l_3\times m},$$

and $r_3=c-C\tilde{y}$, $l_3=2m+1$.

If $\tilde{x} \neq 0$ and $\tilde{y} = 0$, then

$$\eta_{S_2}(\tilde{z}) = \sqrt{\frac{\theta_3^2\mu^2\|c - F\tilde{x}\|_2^2}{\mu^2\|\tilde{x}\|_2^2 + \theta_3^2} + \left\|P_{K_4}^{\perp} d_4\right\|_2^2},$$

where

$$d_4 = \begin{bmatrix} 0 \\ \frac{2\theta_1}{\|\tilde{x}\|_2^2}\tilde{x}^T r_4 \\ \frac{\sqrt{2}\theta_1}{\|\tilde{x}\|_2}\left(I - \tilde{x}\tilde{x}^{\dagger}\right) r_4 \end{bmatrix} \in \mathbb{R}^{l_4}, \quad K_4 = \begin{bmatrix} \lambda I \\ \frac{2\theta_1}{\|\tilde{x}\|_2^2}\tilde{x}^T \\ \frac{\sqrt{2}\theta_1}{\|\tilde{x}\|_2}\left(I - \tilde{x}\tilde{x}^{\dagger}\right) \end{bmatrix} \in \mathbb{R}^{l_4 \times n},$$

and $r_4 = b - A\tilde{x}$, $l_4 = 2n + 1$.

4. Expression of $\eta_{S_3}(\tilde{z})$

In this section we will present the explicit expression of the backward error $\eta_{S_3}(\tilde{z})$ with respect to a computable solution $\tilde{z} = (\tilde{x}^T, \tilde{y}^T)^T$ to the system (1.1) with $E = F$. Its proof is similar to those in Theorems 2.3 and 3.1, which is omitted.

Theorem 4.1. *Let $\tilde{z} = \left(\tilde{x}^T, \tilde{y}^T\right)^T$ be a computed solution of the system (1.1) with $E = F$. Then we have*

$$\eta_{S_3}(\tilde{z}) = \left\|P_{K_5}^{\perp} d_5\right\|_2, \tag{4.1}$$

where

$$d_5 = \begin{bmatrix} 0 \\ \frac{1}{\gamma_1} r_1 \\ \frac{1}{\gamma_2} r_2 \end{bmatrix} \in \mathbb{R}^{l_1}, \quad K_5 = \begin{bmatrix} \theta_2 I \\ \frac{1}{\gamma_1}(\tilde{y}^T \otimes I)\Pi \\ \frac{1}{\gamma_2}\tilde{x}^T \otimes I \end{bmatrix} \in \mathbb{R}^{l_5 \times mn},$$

$$l_5 = mn + m + n, \quad r_1 = b - A\tilde{x} - E^T\tilde{y}, \quad r_2 = c - E\tilde{x} - C\tilde{y},$$

and

$$\gamma_1 = \sqrt{\frac{\|\tilde{x}\|_2^2}{\theta_1^2} + \frac{1}{\lambda^2}}, \quad \gamma_2 = \sqrt{\frac{\|\tilde{y}\|_2^2}{\theta_4^2} + \frac{1}{\mu^2}}.$$

5. A numerical example

To illustrate the application of our formulae, we use the driver *navier_testproblem* (with default parameters) of IFISS [8] package to generate stabilized Q_1-P_0 finite element discretization for the Oseen problem (leaky lid driven cavity) . The generalized saddle-point linear system reads [9]

$$Mz = \begin{bmatrix} A & E^T \\ E & -\frac{1}{\nu}C \end{bmatrix} z = p, \tag{5.1}$$

where $A \neq A^T$, $C = C^T \neq 0$, and $\nu > 0$ represents the viscosity. Following the conventional way, we drop the first row of E to assure the nonsingularity of M.

In the following numerical tests, the right-hand side p in (5.1) is chosen such that the exact solution is all ones. The linear system (5.1) is solved by the Gaussian elimination method with partial pivoting

[11] (GEPP) and the unpreconditioned GMRES method [17], respectively. The initial guess of GMRES is the zero vector and the stopping criterion is $\|r_k\|_2/\|r_0\|_2 \leq 10^{-15}$, where r_k is the residual vector at the kth iteration. For different values of viscosity ν, Table 5.1 reports backward errors of $\eta(\tilde{z})$, $\eta_{S_1}(\tilde{z})$ and $\eta_{S_3}(\tilde{z})$ with respect to approximate solutions $\tilde{z}$, which show that GEPP and GMRES methods are backward stable and strongly stable for solving the system (5.1).

Table 5.1
Backward errors of approximated solutions of system (5.1) on 8×8 grids ($n = 162$, $m = 64 - 1$).

Method	ν	$\|M\tilde{z} - p\|_2/\|p\|_2$	$\eta(\tilde{z})$ in (1.3)	$\eta_{S_1}(\tilde{z})$ in (2.1)	$\eta_{S_3}(\tilde{z})$ in (4.1)
	0.01	1.71e–15	3.07e–17	1.2204e–16	1.1711e–16
	0.10	2.27e–16	1.27e–17	4.7148e–17	4.2719e–17
GEPP	1	5.52e–16	1.44e–17	2.5050e–17	2.5030e–17
	10	6.63e–16	1.27e–17	2.2780e–17	2.2779e–17
	100	6.39e–16	1.21e–17	2.2207e–17	2.2207e–17
	0.01	6.86e–16	4.96e–17	1.8564e–16	1.7301e–16
	0.10	4.83e–16	5.15e–17	1.1918e–16	1.1058e–16
GMRES	1	6.86e–16	4.79e–17	1.4773e–16	1.4754e–16
	10	7.78e–16	4.24e–17	1.2147e–15	1.2147e–15
	100	9.69e–16	4.31e–17	9.0803e–15	9.0803e–15

6. Concluding remarks

In this paper we derive computable expressions for three kinds of generalized saddle point systems where the (2,2)-block is not zero and has perturbations. Our techniques are different from the one described in [20,22], where they first gave computable formula of partial backward errors (see [21,20]), then obtained the expressions of backward errors by using the minimum value of the positive definite quadratic form. However, using the techniques in [21,20], we cannot derive an explicit expression of the backward errors. In this paper, transforming the optimal problems $\eta_{S_1}(\tilde{z})$ and $\eta_{S_2}(\tilde{z})$ into the least squares problems (2.17) and (3.15), respectively, the explicit expressions of the backward error $\eta_{S_i}(\tilde{z})$, $i = 1, 2, 3$, are given by solving the corresponding least squares problem.

Acknowledgment

The authors would like to thank the anonymous referees for their valuable comments.

References

[1] G. Bao, W. Sun, A fast algorithm for the electromagnetic scatting from a large cavity, SIAM J. Sci. Comput. 27 (2005) 553–574.
[2] M. Benzi, G.H. Golub, J. Liesen, Numerical solutions for saddle point problems, Acta Numer. 14 (2005) 1–137.
[3] M.A. Botchev, G.H. Golub, A class of nonsymmetric preconditioners for saddle point problems, SIAM J. Matrix Anal. Appl. 27 (2006) 1125–1149.
[4] J.R. Bunch, The weak and strong stability of algorithms in numerical linear algebra, Linear Algebra Appl. 88/89 (1987) 49–66.
[5] J.R. Bunch, J.W. Demmel, C.F. Van Loan, The strong stability of algorithms for solving symmetric linear systems, SIAM J. Matrix Anal. Appl. 10 (1989) 494–499.
[6] X.S. Chen, W. Li, Structured backward errors for a class of linear systems, Math. Numer. Sin. 29 (2007) 433–438. (in Chinese).
[7] X. Chen, K. Hashimoto, Numerical validation of solutions of saddle point matrix equations, Numer. Linear Algebra Appl. 10 (2003) 661–672.
[8] H.C. Elman, A. Ramage, D.J. Silvester, Algorithm 866: IFISS, a Matlab toolbox for modelling incompressible flow, ACM Trans. Math. Software (TOMS) 33 (2007) 2–14.
[9] H.C. Elman, D.J. Silvester, A.J. Wathen, Finite Elements and Fast Iterative Solvers with Applications in Incompressible Fluid Dynamics, Oxford University Press, Oxford, UK, 2005.
[10] A. Graham, Kronecker Products and Matrix Calculus with Applications, John Wiley, New York, 1981.

[11] G.H. Golub, C. Van Loan, Matrix Computations, third ed., John Hopkins University Press, Baltimore, London, 1996.
[12] N.J. Higham, Accuracy and Stability of Numerical Algorithms, second ed., SIAM, Philadelphia, 2002.
[13] T. Kimura, X. Chen, Validated solution of saddle point linear systems, SIAM J. Matrix Anal. Appl. 30 (2009) 1697–1708.
[14] X. Li, X. Liu, Structured backward errors for structured KKT systems, J. Comput. Math. 22 (2004) 605–610.
[15] M. Nikolova, M.K. Ng, S.Q. Zhang, W.K. Ching, Efficient reconstruction of piecewise constants images using nonsmooth nonconvex minimization, SIAM J. Imaging Sci. 1 (2008) 2–25.
[16] J.L. Rigal, J. Gaches, On the compatibility of a given solution with data of a linear system, J. Assoc. Comput. Mach. 14 (1967) 543–548.
[17] Y. Saad, M.H. Schultz, GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 7 (1986) 856–869.
[18] J.G. Sun, Backward perturbation analysis of certain characteristic subspaces, Numer. Math. 65 (1993) 357–382.
[19] J.G. Sun, Optimal backward perturbation bounds for linear systems and linear least squares problems, UMINF, 96.15, Department of Computing Science, Umeå University, 1996, ISSN 0348-0542.
[20] J.G. Sun, Structured backward errors for KKT systems, Linear Algebra Appl. 288 (1999) 75–88.
[21] J.G. Sun, A note on backward errors for structured linear systems, Numer. Linear Algebra Appl. 12 (2005) 585–603.
[22] H. Xiang, Y.M. Wei, On normwise structured backward errors for saddle point systems, SIAM J. Matrix Anal. Appl. 29 (2007) 838–849.